

1 Genomic epidemiology reveals transmission patterns and 2 dynamics of SARS-CoV-2 in Aotearoa New Zealand

3
4 Jemma L Geoghegan^{1,2}, Xiaoyun Ren², Matthew Storey², James Hadfield³, Lauren Jelley², Sarah
5 Jefferies², Jill Sherwood², Shevaun Paine², Sue Huang², Jordan Douglas⁴, Fábio K Mendes⁴,
6 Andrew Sporle^{5,6}, Michael G Baker⁷, David R Murdoch⁸, Nigel French⁹, Colin R Simpson^{10,11}, David
7 Welch⁴, Alexei J Drummond⁴, Edward C Holmes¹², Sebastián Duchêne¹³, Joep de Ligt²

8
9 ¹Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand.

10 ²Institute of Environmental Science and Research, Wellington, New Zealand.

11 ³Fred Hutchinson Cancer Research Centre, Seattle, Washington, USA.

12 ⁴Centre for Computational Evolution, School of Computer Science, University of Auckland,
13 Auckland, New Zealand.

14 ⁵Department of Statistics, University of Auckland, New Zealand.

15 ⁶McDonaldSporle Ltd., Auckland, New Zealand.

16 ⁷Department of Public Health, University of Otago, Wellington, New Zealand.

17 ⁸Department of Pathology and Biomedical Science, University of Otago, Christchurch, New
18 Zealand.

19 ⁹School of Veterinary Science, Massey University, Palmerston North, New Zealand.

20 ¹⁰School of Health, Faculty of Health, Victoria University of Wellington, Wellington, New Zealand.

21 ¹¹Usher Institute, University of Edinburgh, Edinburgh, United Kingdom.

22 ¹²Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental
23 Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales,
24 Australia.

25 ¹³Department of Microbiology and Immunology, The University of Melbourne at The Peter Doherty
26 Institute for Infection and Immunity, Melbourne, Victoria, Australia.

27 Author for correspondence: jemma.geoghegan@otago.ac.nz

28 Keywords: SARS-CoV-2; COVID-19; coronavirus; genomics; phylodynamics; phylogenetics; virus
29 evolution; infectious disease; New Zealand

This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

30 Abstract

31 New Zealand, a geographically remote Pacific island with easily sealable borders, implemented a
32 nation-wide lockdown of all non-essential services to curb the spread of COVID-19. New Zealand
33 has now effectively eliminated the virus, with low numbers of new cases limited to new arrivals in
34 managed quarantine facilities at the border. Here, we generated 649 SARS-CoV-2 genome
35 sequences from infected patients in New Zealand with samples collected between 26 February
36 and 22 May 2020, representing 56% of all confirmed cases in this time period. Despite its
37 remoteness, the viruses imported into New Zealand represented nearly all of the genomic
38 diversity sequenced from the global virus population. The proportion of D614G variants in the
39 virus spike protein increased over time due to an increase in their importation frequency, rather
40 than selection within New Zealand. These data also helped to quantify the effectiveness of public
41 health interventions. For example, the effective reproductive number, R_e , of New Zealand's largest
42 cluster decreased from 7 to 0.2 within the first week of lockdown. Similarly, only 19% of virus
43 introductions into New Zealand resulted in a transmission lineage of more than one additional
44 case. Most of the cases that resulted in a transmission lineage originated from North America,
45 rather than from Asia where the virus first emerged or from the nearest geographical neighbour,
46 Australia. Genomic data also helped link more infections to a major transmission cluster than
47 through epidemiological data alone, providing probable sources of infections for cases in which
48 the source was unclear. Overall, these results demonstrate the utility of genomic pathogen
49 surveillance to inform public health and disease mitigation.

50

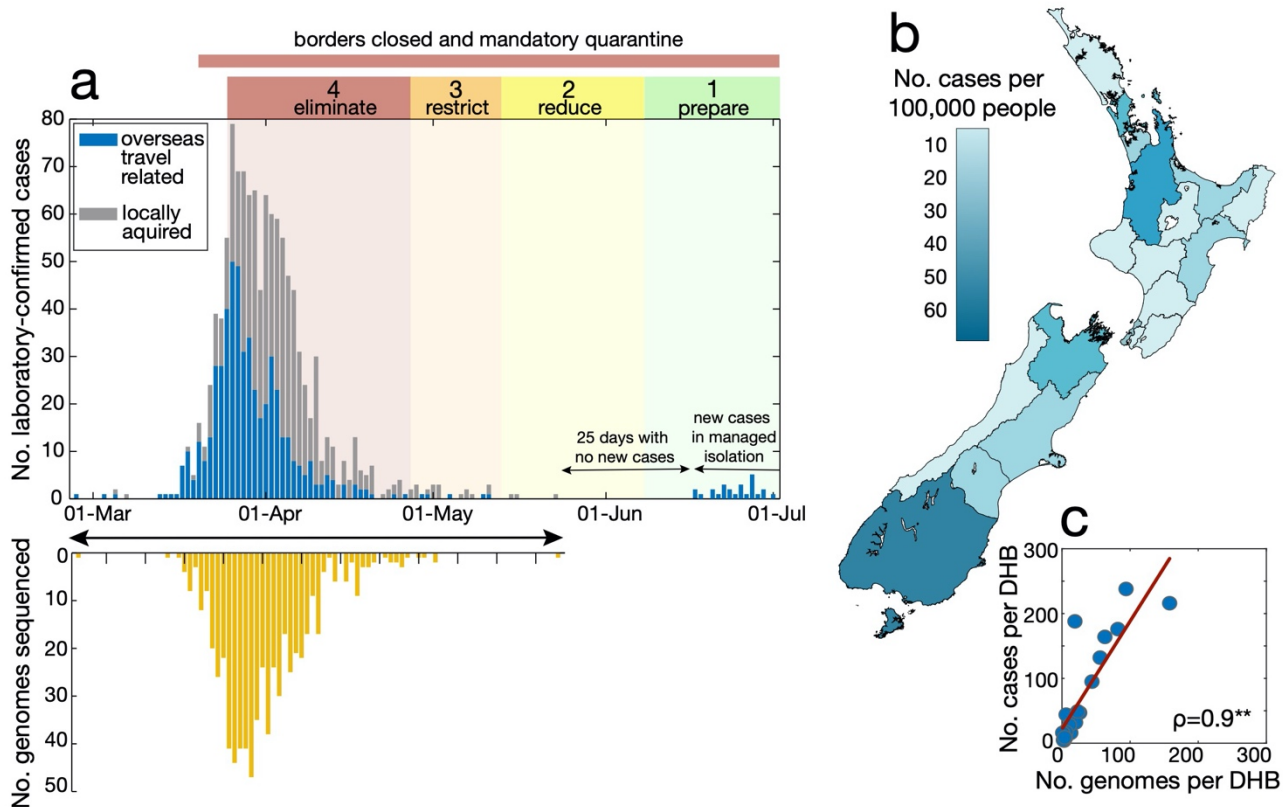
51 Main Text

52 New Zealand is one of a handful of countries that aimed to eliminate coronavirus disease 19
53 (COVID-19). The disease was declared a global pandemic by the World Health Organisation
54 (WHO) on 11 March 2020. The causative virus, severe acute respiratory syndrome coronavirus 2
55 (SARS-CoV-2)¹, was first identified and reported in China in late December 2019, and is the
56 seventh coronavirus known to infect humans, likely arising through zoonotic transmission from
57 wildlife². Because of its relatively high case fatality rate³⁻⁵, and virus transmission from
58 asymptomatic or pre-symptomatic individuals^{6,7}, SARS-CoV-2 presents a significant public health
59 challenge. Due to its high rate of transmission, morbidity and mortality, SARS-CoV-2 has resulted
60 in world-wide lockdowns, economic collapses and led to healthcare systems being overrun.

61 Since the publication of the first SARS-CoV-2 genome on 10 January 2020⁸, there has been a
62 substantial global effort to contribute and share genomic data to inform local and international
63 communities about key aspects of the pandemic⁹. Analyses of genomic data have played an
64 important role in tracking the epidemiology and evolution of the virus, often doing so in real time¹⁰,

65 and leading to a greater understanding of COVID-19 outbreaks globally¹¹⁻¹⁵.

66 New Zealand reported its first case on 26 February 2020 and within a month implemented a
67 stringent, country-wide lockdown of all non-essential services. To investigate the origins, time-
68 scale and duration of virus introductions into New Zealand, the extent and pattern of viral spread
69 across the country, and to quantify the effectiveness of intervention measures, we generated
70 whole genome sequences from 56% of all documented SARS-CoV-2 cases from New Zealand
71 and combined these with detailed epidemiological data.



72
73 **Figure 1.** (a) Number of laboratory-confirmed cases by reported date, both locally acquired (grey)
74 and linked to overseas travel (blue) in New Zealand, highlighting the timing of public health alert
75 levels 1-4 ('eliminate', 'restrict', 'reduce', 'prepare') and national border closures. The number of
76 genomes sequenced in this study is shown over time. (b) Map of New Zealand's District Health
77 Boards shaded by the incidence of laboratory-confirmed cases of COVID-19 per 100,000 people.
78 (c) Number of laboratory-confirmed cases per District Health Board (DHB) versus number of
79 genomes sequenced, indicating Spearman's ρ , where asterisks indicate statistical significance
80 ($p < 0.0001$).

81

82 **Table 1.** Demographic data for confirmed (n=1178) and probable (n=350) cases of SARS-CoV-2
83 in New Zealand between 26 February and 1 July 2020. The percentage of genomes sequenced in
84 each category is shown.

85

Age group	Number of cases	Deceased	Percentage of genomes in data set
0 to 9	37	0	6%
10 to 19	122	0	38%
20 to 29	365	0	45%
30 to 39	238	0	39%
40 to 49	221	0	42%
50 to 59	248	0	44%
60 to 69	180	3	45%
70 to 79	78	7	45%
80 to 89	30	7	50%
90+	9	5	56%

Gender	Number of cases	Percentage of cases	Percentage of genomes in data set
Female	848	55%	42%
Male	680	45%	41%

Ethnicity	Number of cases	Percentage of cases	Percentage of genomes in data set
European or other	1067	70%	46%
Asian	210	14%	27%
Māori	130	9%	42%
Pacific peoples	81	5%	35%
Middle Eastern / Latin American / African	33	2%	42%
Unknown	7	0.50%	86%

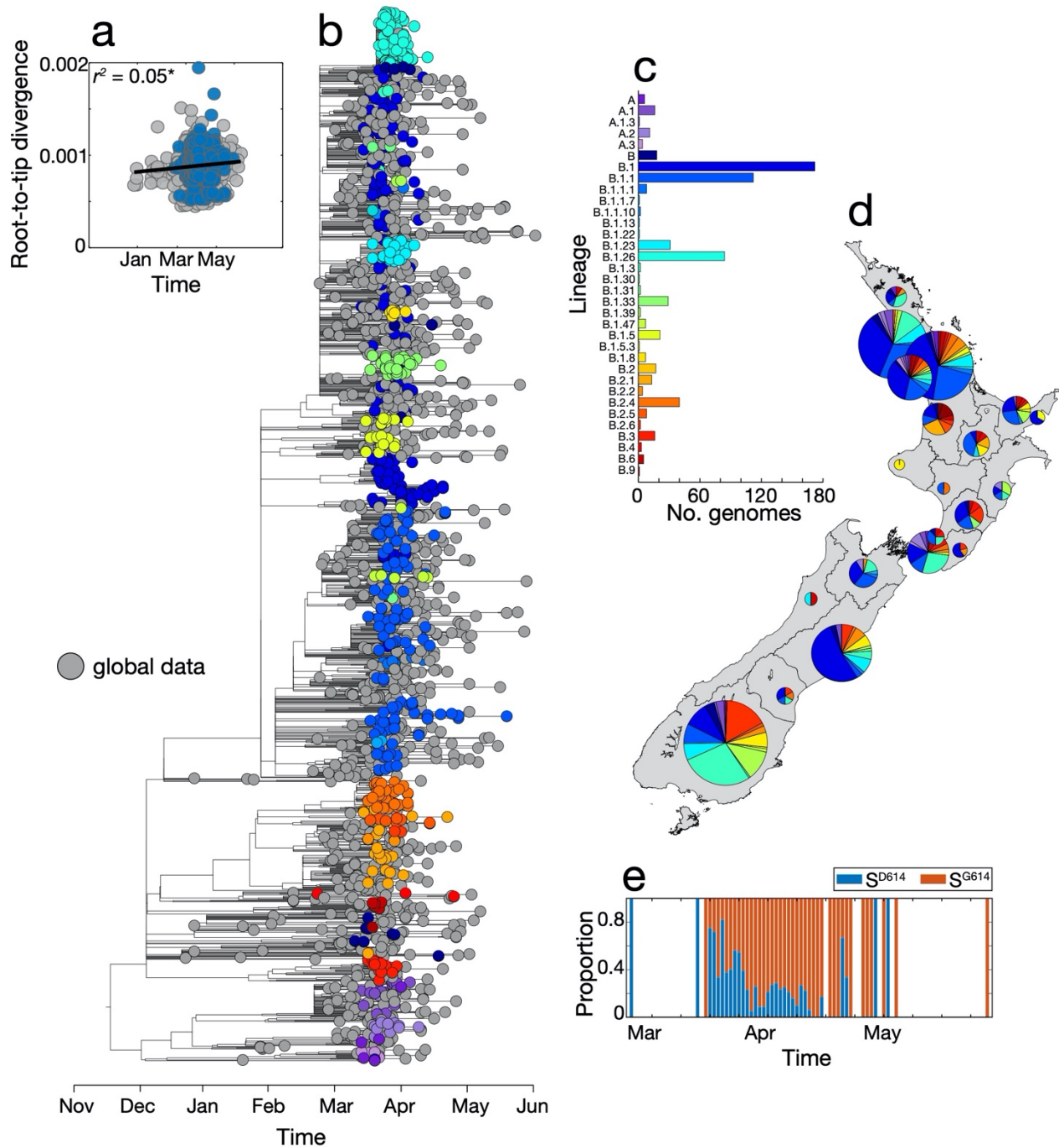
Transmission type	Number of cases	Percentage of cases	Percentage of genomes in data set
Imported cases	572	37%	48%
Locally-acquired cases	956	63%	39%

86
87
88
89

90 Between 26 of February and 1 July 2020 there were a total of 1,178 laboratory-confirmed cases
91 and a further 350 probable cases of SARS-CoV-2 in New Zealand (a probable case is defined as
92 a person who has returned a negative laboratory result or could not be tested, but the medical
93 officer of health has assigned the case classification based on exposure history and clinical
94 symptoms). Of these combined laboratory-confirmed and probable cases, 55% were female and
95 45% were male, with the highest proportion of cases in the 20-29 age group (Table 1). Many
96 cases were linked to overseas travel (37%). Geographic locations in New Zealand with the highest
97 number of reported cases did not necessarily reflect the human population size or density in that
98 region, with the highest incidence reported in the Southern District Health Board (DHB) region
99 rather than in highly populated cities (Figure 1). The number of laboratory-confirmed cases
100 peaked on 26 March 2020, the day after New Zealand instigated an Alert Level 4 lockdown – the
101 most stringent level, ceasing all non-essential services and stipulating that the entire population
102 self-isolate (Figure 1). From 23 May 2020, New Zealand experienced 25 consecutive days with no
103 new reported cases until 16 June, when new infections, linked to overseas travel, were
104 diagnosed. All subsequent new cases have been from patients in managed quarantine facilities.

105 We sequenced a total of 649 virus genomes from samples taken between 26 February (first
106 reported case) and 22 May 2020 (to date, the last confirmed case unassociated with managed
107 quarantine facilities). This represented 56% of all New Zealand’s confirmed cases sampled during
108 that period. The data generated originated from the 20 DHBs from across New Zealand. DHBs
109 submitted between 0.1% and 81% of their positive samples to the Institute of Environmental
110 Science and Research (ESR), Wellington, for sequencing. Despite this disparity, a strong
111 nationwide spatial representation was achieved (Figure 1).

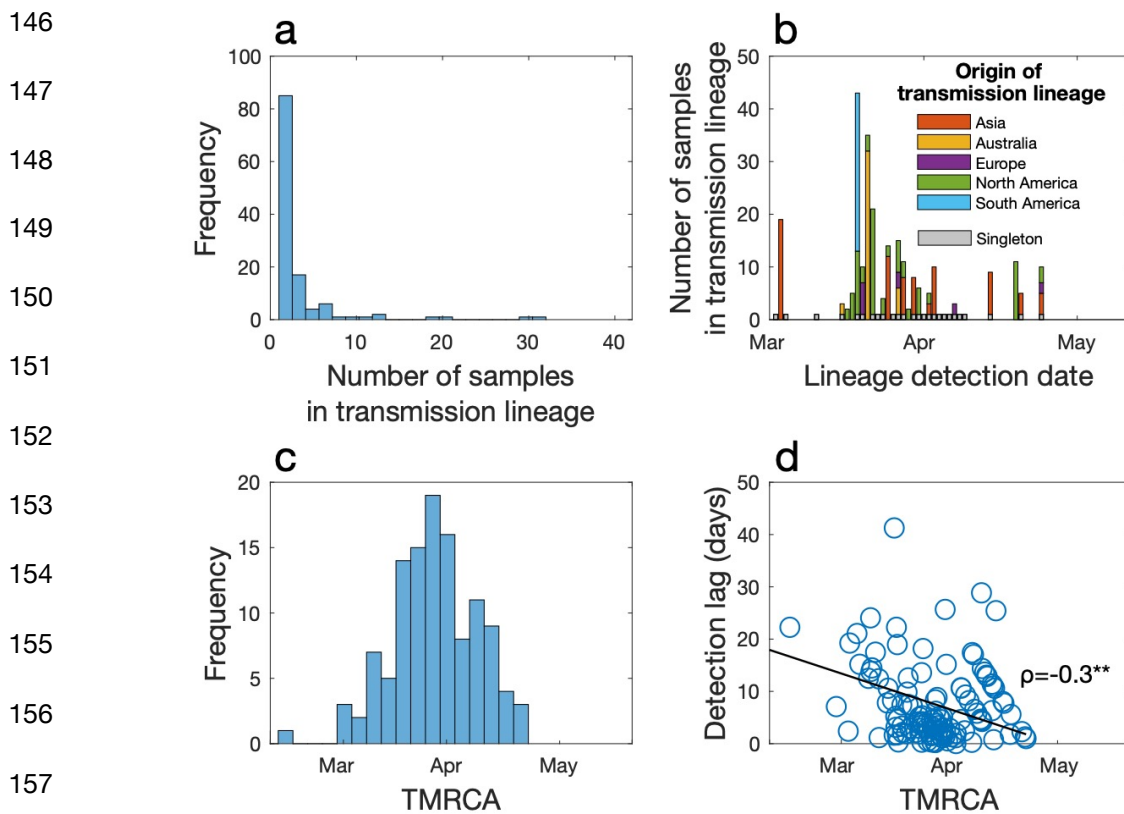
112 Notably, the genomic diversity of SARS-CoV-2 sequences sampled in New Zealand represented
113 nearly all of the genomic diversity present in the global viral population, with nine second-level A
114 and B lineages from a recently proposed global SARS-CoV-2 genomic nomenclature¹⁶ identified.
115 This high degree of genomic diversity was observed throughout the country (Figure 2). The SARS-
116 CoV-2 genomes sampled in New Zealand comprised 24% aspartic acid (S^{D614}) and 73% glycine
117 (S^{G614}) at residue 614 in the spike protein (Figure 2). Preliminary studies suggest that the D614G
118 mutation can enhance viral infectivity in cell culture¹⁷. Nevertheless, it is noteworthy that the
119 increase in glycine in New Zealand samples is due to multiple importation events of this variant
120 rather than selection for this mutation within New Zealand. We also inferred a weak yet significant
121 temporal signal in the data, reflecting the low mutation rate of SARS-CoV-2, which is consistent
122 with findings reported elsewhere (Figure 2).



123

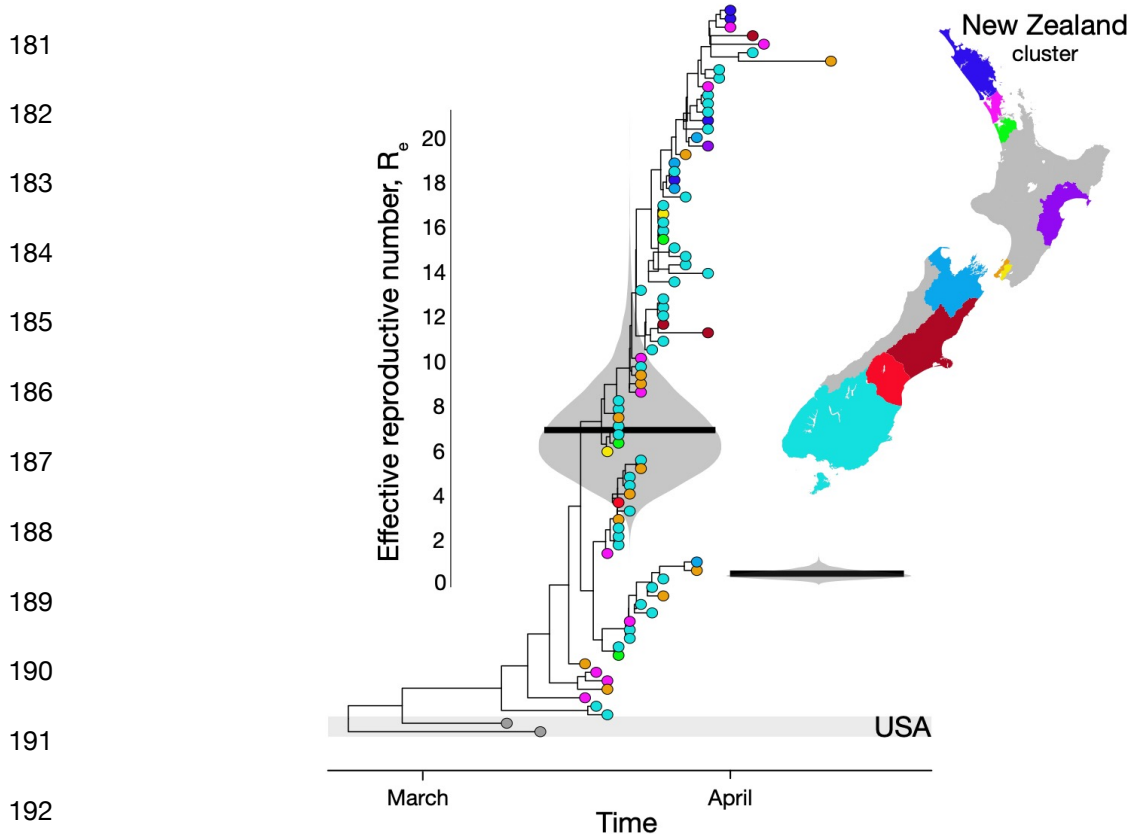
124 **Figure 2.** (a) Root-to-tip regression analysis of New Zealand (blue) and global (grey) SARS-CoV-2
 125 sequences, with the determination coefficient, r^2 (an asterisk indicates statistical significance;
 126 $p < 0.05$). (b) Maximum-likelihood time-scaled phylogenetic analysis of 649 viruses sampled from
 127 New Zealand (coloured circles) on a background of 1000 randomly subsampled viruses from the
 128 globally available data (grey circles). Viruses sampled from New Zealand are colour-coded
 129 according to their genomic lineage¹⁶. (c) The number of SARS-CoV-2 genomes sampled in New
 130 Zealand within each lineage¹⁶. (d) The sampling location and proportion of SARS-CoV-2 genomes
 131 sampled from each viral genomic lineage is shown on the map of New Zealand. (e) The frequency
 132 of D (blue) and G (red) amino acids at residue 614 on the spike protein over time.

133 Despite the small size of the New Zealand outbreak, there were 277 separate introductions of the
134 virus out of the 649 cases considered. Of these, we estimated that 24% (95% CI: 23-30) led to
135 only one other secondary case (i.e. singleton) while just 19% (95% CI: 15-20) of these introduced
136 cases led to ongoing transmission, forming a transmission lineage (i.e. onward transmission to
137 more than one individual; Figure 3). The remainder (57%) did not lead to a transmission event.
138 New Zealand transmission lineages most often originated in North America, rather than in Asia
139 where the virus first emerged, likely reflecting the high prevalence of the virus in North America
140 during the sampling period. By examining the time of the most recent common ancestor, or
141 TMRCA, of the samples, we found no evidence that the virus was circulating in New Zealand
142 before the first reported case on 26 February. Finally, we found that detection was more efficient
143 (i.e. fewer cases were missed) later in the epidemic in that the detection lag (the duration of time
144 from the first inferred transmission event to the first detected case) declined with the age of
145 transmission lineages (as measured by the time between the present and the TMRCA; Figure 3).



158 **Figure 3.** (a) Frequency of transmission lineage size. (b) The number of samples in each
159 transmission lineage as a function of the date at which the transmission lineage was sampled,
160 coloured by the likely origin of each lineage (inferred from epidemiological data). Importation
161 events that led to only one additional case (singletons) are shown in grey over time. (c) Frequency
162 of TMRCA (the time of the most recent common ancestor) of importation events over time. (d) The
163 difference between the TMRCA and the date as which a transmission lineage was detected (i.e.
164 detection lag) as a function of TMRCA. Spearman's ρ indicates a significant negative relationship
165 ($p < 0.01$).

166 The largest clusters in New Zealand were often associated with social gatherings such as
167 weddings, hospitality and conferences¹⁸. The largest cluster, which comprised lineage B.1.26,
168 most likely originated in the USA according to epidemiological data, and significant local
169 transmission in New Zealand was probably initiated by a superspreading event at a wedding in
170 Southern DHB (geographically the most southern DHB) prior to lockdown. Examining the rate of
171 transmission of this cluster enables us to quantify the effectiveness of the lockdown. Its effective
172 reproductive number, R_e , decreased over time from 7 at the beginning of the outbreak (95%
173 credible interval, CI: 3.7-10.7) to 0.2 (95% CI: 0.1-0.4) by the end of March (Figure 4). The
174 sampling proportion of this cluster, a key parameter of the model, had a mean of 0.75 (95% CI:
175 0.4-1), suggesting sequencing captured the majority of cases in this outbreak. In addition,
176 analysis of genomic data has linked five additional cases to this cluster that were not identified in
177 the initial epidemiological investigation, highlighting the added value of genomic analysis. This
178 cluster, seeded by a single-superspreading event that resulted in New Zealand's largest chain of
179 transmission, illustrates the link between micro-scale transmission to nation-wide spread (Figure
180 4).



193 **Figure 4.** Maximum clade credibility phylogenetic tree of New Zealand's largest cluster with an
194 infection that most likely originated in the USA. Estimates of the effective reproductive number,
195 R_e , are shown in violin plots superimposed onto the tree, grouping the New Zealand samples into
196 two time-intervals as determined by the model. Black horizontal lines indicate the mean R_e . Tips
197 are coloured by the reporting District Health Board and their location shown on the map.

198 The dramatic decrease in R_e of this large cluster coupled with the relatively low number of virus
199 introductions that resulted in a transmission lineage suggests that implementing a strict and early
200 lockdown in New Zealand rapidly reduced multiple chains of virus transmission. As New Zealand
201 continues to remain free of COVID-19 community transmission, but with positive cases still
202 detected amongst individuals quarantined at the border reflecting high virus incidence in other
203 localities, it is imperative that ongoing genomic surveillance is an integral part of the national
204 response to monitor any re-emergence of the virus, particularly when border restrictions might
205 eventually be eased.

206

207 Methods

208 **Ethics statement.** Nasopharyngeal samples testing positive for SARS-CoV-2 by real-time
209 polymerase chain reaction (RT-PCR) were obtained from public health medical diagnostics
210 laboratories located throughout New Zealand. All samples were de-identified before receipt by the
211 researchers. Under contract for the Ministry of Health, ESR has approval to conduct genomic
212 sequencing for surveillance of notifiable diseases.

213 **Genomic sequencing of SARS-CoV-2.** A total of 733 laboratory-confirmed samples of SARS-
214 CoV-2 were received by ESR for whole genome sequencing. Viral extracts were prepared from
215 respiratory tract samples where SARS-CoV-2 was detected by RT-PCR using WHO
216 recommended primers and probes targeting the E and N gene. Extracted RNA from SARS-CoV-2
217 positive samples were subject to whole genome sequencing following the ARTIC network
218 protocol (V1 and V3) and the New South Wales (NSW) primer set¹⁵.

219 Briefly, three different tiling amplicon designs were used to amplify viral cDNA prepared with
220 SuperScript IV. Sequence libraries were then constructed using Illumina Nextera XT for the NSW
221 primer set or the Oxford Nanopore ligation sequencing kit for the ARTIC protocol. Libraries were
222 sequenced using Illumina NextSeq chemistry or R9.4.1 MinION flow cells, respectively. Near-
223 complete (>90% recovered) viral genomes were subsequently assembled through reference
224 mapping. Steps included in the pipeline are described in detail online ([https://github.com/ESR-
225 NZ/NZ_SARS-CoV-2_genomics](https://github.com/ESR-NZ/NZ_SARS-CoV-2_genomics)).

226 The reads generated with Nanopore sequencing using ARTIC primer sets (V1 and V3) were
227 mapped and assembled using the ARTIC bioinformatics medaka pipeline (v 1.1.0)¹⁹. For the NSW
228 primer set, raw reads were quality and adapter trimmed using trimmomatic (v 0.36)²⁰. Trimmed
229 paired reads were mapped to a reference using the Burrows-Wheeler Alignment tool²¹. Primer
230 sequences were masked using iVar (v 1.2)²². Duplicated reads were marked using Picard (v
231 2.10.10)²³ and not used for SNP calling or depth calculation. Single nucleotide polymorphisms

232 (SNPs) were called using bcftools mpileup (v 1.9)²⁴. SNPs were quality trimmed using vcflib (v
233 1.0.0)²⁵ requiring 20x depth and overall quality of 30. Positions that were less than 20x were
234 masked to N in the final consensus genome. Positions with an alternative allele frequency
235 between 20% to 79% were also masked to N. In total, 649 sequences passed our quality control
236 (BioProject: PRJNA648792; a list of genomes and their sequencing methods are provided in
237 Supplementary Table 1).

238 **Phylogenetic analysis of SARS-CoV-2.**

239 SARS-CoV-2 sequences from New Zealand, together with 1,000 genomes uniformly sampled at
240 random from the global population from the ~50,000 available sequences from GISAID²⁶ (June
241 2020), were aligned using MAFFT(v 7)²⁷ using the FFT-NS-2 algorithm. A maximum likelihood
242 phylogenetic tree was estimated using IQ-TREE (v 1.6.8)²⁸, utilising the Hasegawa-Kishino-Yano
243 (HKY+ Γ)²⁹ nucleotide substitution model with a gamma distributed rate variation among sites (the
244 best fit model was determined by ModelFinder³⁰), and branch support assessment using the
245 ultrafast bootstrap method³¹. We regressed root-to-tip genetic divergence against sampling dates
246 to investigate the evolutionary tempo of our SARS-CoV-2 samples using TempEst (v 1.5.3)³².
247 Lineages were assigned according to the proposed nomenclature¹⁶ using pangolin
248 (<https://github.com/hCoV-2019/pangolin>). To depict virus evolution in time, we used Least
249 Squares Dating³³ to estimate a time-scaled phylogenetic tree using the day of sampling.

250 With the full set of New Zealand sequences, we used a time-aware coalescent Bayesian
251 exponential growth model available in BEAST (v 1.10.4)³⁴. The HKY+ Γ model of nucleotide
252 substitution was again used along with a strict molecular clock. Because the data did not display
253 a strong temporal signal, we used an informative prior reflecting recent estimates for the
254 substitution rate of SARS-CoV-2³⁵. The clock rate had a Γ prior distribution as a prior with a mean
255 of 0.8×10^{-3} subs/site/year and standard deviation of 5×10^{-4} (parameterised using the shape and
256 rate of the Γ distribution). Parameters were estimated using Bayesian Markov Chain Monte Carlo
257 (MCMC) framework, with 2×10^8 steps-long chains, sampling every 1×10^5 steps and removing
258 the initial 10% as burn-in. Sufficient sampling was assessed using Tracer (v 1.7.1)³⁶, by verifying
259 that every parameter had effective sampling sizes above 200. Virus sequences were annotated as
260 ‘imported’ (including country of origin) or ‘locally acquired’, according to epidemiological data
261 provided by EpiSurv³⁷. From a set of 1,000 posterior trees, we estimated a number of statistics
262 using NELSI³⁸. We determined the number of introductions of the virus into New Zealand as well
263 as the changing number of local transmission lineages through time, with the latter defined as two
264 or more New Zealand SARS-CoV-2 cases that descend from a shared introduction event of the
265 virus into New Zealand³⁹. Importation events that led to only a single case rather than a
266 transmission lineage are referred to as ‘singletons’. For each transmission lineage and singleton,
267 we inferred the TMRCA.

268 To estimate R_e through time we analysed New Zealand sequences from the clade identified to be
269 associated with a wedding. We used a Bayesian birth-death skyline model using BEAST (v 2.5)⁴⁰,
270 estimating R_e for two time-intervals, as determined by the model, and with the same parameter
271 settings as above. We assumed an infectious period of 10 days, which is consistent with global
272 epidemiological estimates⁴¹.

273

274 Online Supplementary Material

275 **Supplementary Table 1.** A list of genomes and which amplification and sequencing method was
276 used in for each case.

277

278 Acknowledgements

279 This work was funded by the Ministry of Health of New Zealand, New Zealand Ministry of
280 Business, Innovation and Employment COVID-19 Innovation Acceleration Fund (CIAF-0470), ESR
281 Strategic Innovation Fund and the New Zealand Health Research Council (20/1018). We thank the
282 ATRIC network for making their protocols and tools openly available and specifically Josh Quick
283 for sending the initial V1 and V3 amplification primers. We thank Genomics Aotearoa for their
284 support. We thank the diagnostic laboratories that performed the initial RT-PCRs and referred
285 samples for sequencing as well as the public health units for providing epidemiological data. We
286 thank the Nextstrain team for their support and timely global and local analysis. We thank all
287 those who have contributed SARS-CoV-2 sequences to GenBank and GISAID databases.

288

289 References

- 290 1 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China.
291 *Nature* **579**, 265–269, doi:10.1038/s41586-020-2008-3 (2020).
- 292 2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat
293 origin. *Nature* **579**, 270–273, doi:10.1038/s41586-020-2012-7 (2020).
- 294 3 Wu, J. T. *et al.* Estimating clinical severity of COVID-19 from the transmission dynamics in
295 Wuhan, China. *Nat. Med.* **26**, 506–510, doi:10.1038/s41591-020-0822-7 (2020).
- 296 4 Russell, T. W. *et al.* Estimating the infection and case fatality ratio for coronavirus disease
297 (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise
298 ship, February 2020. *Euro Surveill.* **25**, doi:10.2807/1560-7917.Es.2020.25.12.2000256
299 (2020).
- 300 5 Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based
301 analysis. *Lancet Infect. Dis.* **20**, 669–677, doi:10.1016/s1473-3099(20)30243-7 (2020).
- 302 6 Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with
303 digital contact tracing. *Science* **368**, eabb6936, doi:10.1126/science.abb6936 (2020).
- 304 7 Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic
305 proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess
306 cruise ship, Yokohama, Japan, 2020. *Euro surveill.* **25**, 2000180, doi:10.2807/1560-
307 7917.ES.2020.25.10.2000180 (2020).
- 308 8 Holmes, E. C. Novel 2019 coronavirus genome, [https://virological.org/t/novel-2019-](https://virological.org/t/novel-2019-coronavirus-genome/319)
309 [coronavirus-genome/319](https://virological.org/t/novel-2019-coronavirus-genome/319) (2020).
- 310 9 Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.*
311 **4**, 10–19, doi:10.1038/s41564-018-0296-2 (2019).
- 312 10 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinform* **34**, 4121–
313 4123, doi:10.1093/bioinformatics/bty407 (2018).
- 314 11 Candido, D. d. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*
315 eabd216, 10.1126/science.abd2161 (2020).
- 316 12 Filipe, A. D. S. *et al.* Genomic epidemiology of SARS-CoV-2 spread in Scotland highlights
317 the role of European travel in COVID-19 emergence. *medRxiv*, 2020.2006.2008.20124834,
318 doi:10.1101/2020.06.08.20124834 (2020).
- 319 13 Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics.
320 *medRxiv*, 2020.2005.2012.20099929, doi:10.1101/2020.05.12.20099929 (2020).
- 321 14 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv*,
322 2020.2004.2002.20051417, doi:10.1101/2020.04.02.20051417 (2020).
- 323 15 Eden, J.S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran.
324 *Virus Evol.* **6**, doi:10.1093/ve/veaa027 (2020).

- 325 16 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic
326 epidemiology. *bioRxiv*, 2020.2004.2017.046086, doi:10.1101/2020.04.17.046086 (2020).
- 327 17 Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1
328 shedding and increases infectivity. *bioRxiv*, 2020.2006.2012.148726,
329 doi:10.1101/2020.06.12.148726 (2020).
- 330 18 Leclerc, Q. J. *et al.* What settings have been linked to SARS-CoV-2 transmission clusters?
331 *Wellcome Open Res.* **5**, 83, doi:10.12688/wellcomeopenres.15889.2 (2020).
- 332 19 Loman, N. R., W; Rambaut, A. nCoV-2019 novel coronavirus bioinformatics protocol,
333 <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html> (2020).
- 334 20 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
335 sequence data. *Bioinform.* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
- 336 21 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
337 transform. *Bioinform.* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
- 338 22 Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring
339 intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8,
340 doi:10.1186/s13059-018-1618-7 (2019).
- 341 23 Picard Toolkit. Broad Institute, <http://broadinstitute.github.io/picard/> (2019).
- 342 24 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
343 and population genetical parameter estimation from sequencing data. *Bioinform.* **27**,
344 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
- 345 25 Garrison, E. Vcflib, a simple C++ library for parsing and manipulating VCF files.
346 <https://github.com/vcflib/vcflib> (2016).
- 347 26 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative
348 contribution to global health. *Global Challenges* **1**, 33-46, doi:10.1002/gch2.1018 (2017).
- 349 27 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
350 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780,
351 doi:10.1093/molbev/mst010 (2013).
- 352 28 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
353 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**,
354 268-274, doi:10.1093/molbev/msu300 (2015).
- 355 29 Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the human-ape splitting by a molecular
356 clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174, doi:10.1007/BF02101694 (1985).
- 357 30 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S.
358 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**,
359 587-589, doi:10.1038/nmeth.4285 (2017).

- 360 31 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
361 Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518-522,
362 doi:10.1093/molbev/msx28 (2017).
- 363 32 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure
364 of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**,
365 vew007, doi:10.1093/ve/vew007 (2016).
- 366 33 To, T. H., Jung, M., Lycett, S. & Gascuel, O. Fast dating using least-squares criteria and
367 algorithms. *Syst. Biol.* **65**, 82-97, doi:10.1093/sysbio/syv068 (2016).
- 368 34 Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees.
369 *BMC Evol. Biol.* **7**, 214, doi:10.1186/1471-2148-7-214 (2007).
- 370 35 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal
371 origin of SARS-CoV-2. *Nat. Med.* **26**, 450-452, doi:10.1038/s41591-020-0820-9 (2020).
- 372 36 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior
373 Summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901-904,
374 doi:10.1093/sysbio/syy032 (2018).
- 375 37 EpiSurv: national notifiable disease surveillance database,
376 <https://surv.esr.cri.nz/episurv/index.php>, (2020).
- 377 38 Ho, S. Y., Duchêne, S. & Duchêne, D. Simulating and detecting autocorrelation of
378 molecular evolutionary rates among lineages. *Mol. Ecol. Resour.* **15**, 688-696,
379 doi:10.1111/1755-0998.12320 (2015).
- 380 39 Pybus, O. G. Preliminary analysis of SARS-CoV-2 importation & establishment of UK
381 transmission lineages, [https://virological.org/t/preliminary-analysis-of-sars-cov-2-](https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507)
382 [importation-establishment-of-uk-transmission-lineages/507](https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507) (2020).
- 383 40 Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals
384 temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad.*
385 *Sci. USA* **110**, 228-233, doi:10.1073/pnas.1207965110 (2013).
- 386 41 He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat.*
387 *Med.* **26**, 672-675, doi:10.1038/s41591-020-0869-5 (2020).

388

389